Classer pour Découvrir : une nouvelle méthode d'analyse du comportement de tous les utilisateurs d'un site Web

Doru Tanasa, Brigitte Trousse, Florent Masseglia

Projet AxIS, INRIA Sophia Antipolis,
2004, Route des Lucioles, BP 93, 06902 Sophia Antipolis
{Doru.Tanasa, Brigitte.Trousse, Florent.Masseglia}@sophia.inria.fr,
http://www-sop.inria.fr/axis/

Résumé. L'analyse du comportement des utilisateurs d'un site Web est un domaine riche et complexe. Le grand nombre de méthodes d'extraction de connaissances appliquées aux logs Web, ainsi que la diversité du type de ces méthodes en est une preuve. Cependant, compte tenu de cette complexité, nous posons dans cet article la question suivante : *Est-il possible de combiner des méthodes existantes pour proposer une analyse qui tire profit des résultats de plusieurs spécialités et extraire par exemple des comportements fréquents minoritaires?* Notre étude a donc porté sur une nouvelle approche hybride (issue de la classification neuronale et de la recherche de motifs séquentiels) visant à classer les navigations des utilisateurs d'un site (à l'aide de leurs résumés sémantiques) puis, pour chaque classe de navigations, d'en extraire les comportements fréquents. Notre objectif est 1) de pallier les limites de l'extraction des motifs fréquents par rapport à la quantité de données à traiter et aussi par rapport à la qualité des résultats et 2) de pallier les limites d'une première méthode d'analyse du comportement appelée "Diviser pour Découvrir", que nous avons proposée en 2003. Nous avons mené des expérimentations sur les logs HTTP des sites INRIA. Les résultats obtenus confirment le bien fondé de notre approche vis à vis de l'état de l'art.

1 Introduction

L'analyse du comportement des utilisateurs d'un site Web, également connue sous le nom de Web Usage Mining (WUM), est un domaine de recherche qui consiste à adapter des techniques de fouille de données sur les enregistrements contenus dans les fichiers access logs. Ces fichiers regroupent des informations sur l'adresse IP de la machine, l'URL demandée, la date, et d'autres renseignements concernant la navigation de l'utilisateur. Les techniques de Web Usage Mining s'intéressent à la recherche de motifs (voire des connaissances sur les comportements des utilisateurs d'un site Web) afin d'extraire des relations entre les données stockées [Cooley *et al.*, 1999] [Masseglia *et al.*, 2000] [Mobasher *et al.*, 2002] [Spiliopoulou *et al.*, 1999]. Parmi les méthodes développées, celles consistant à extraire des motifs séquentiels [Agrawal et Srikant, 1995] s'adaptent particulièrement bien au cas des logs. En théorie, l'extraction de motifs séquentiels sur un fichier access log permet d'extraire le type de relations suivant : "*Sur le site de l'IN-*